Spring 2015
Chris Christensen
MAT/CSC 483

Cryptanalysis of the Vigenère Cipher: The Friedman Test

For the Kasiski attack to work, it is necessary for the keyword to be repeated. In fact, what we depended upon was that we had a very long message and relatively short keyword so that the keyword was repeated many times and that when we stripped off the various Caesar cipher alphabets each alphabet contained enough letters to enable us to spot the shift.

In 1922 William Friedman, who is often called the Dean of American Cryptology, published a statistical test that can be used to determine whether a cipher is polyalphabetic or monoalphabetic and for polyalphabet ciphers can estimate the number of alphabets (the length of the keyword for the Vigenère cipher).

William Friedman (1891 – 1969) retired from the National Security Agency in 1955 after 35 years of service with U.S. cryptological activities. Friedman transformed the methods and approaches of cryptology from the traditional into the modern. His wife Elizebeth was also a cryptologist and served at one point with the Coast Guard; she cryptanalyzed messages of the rumrunners. The two of them effectively debunked the theory that Francis Bacon used steganography to conceal and reveal that he was the author of Shakespeare's works (*The Shakespearean Cipher Examined,* 1957).

Friedman's test is a statistical test based upon frequency. One calculation is to determine the **index of coincidence** *I*. (Because Friedman denoted this number by the Greek letter kappa κ, it is sometimes called the Kappa Test.) *I* varies between approximately 0.038 and 0.065. A value of *I* near 0.065 would indicate that a monoalphabetic cipher (like a simple substitution cipher, a Caesar cipher, a multiplicative cipher, an affine cipher, or a keyword cipher) was used, and a value of *I* near 0.038 would indicate that a polyalphabetic cipher (like the Vigenère cipher) was used.

The other calculation is an **approximation to the length of the keyword** *l*.

Index of coincidence

The index of coincidence is sometimes called the repeat rate. Friedman had noticed that when drawing two ciphertext letters at random the probability of drawing "doubles," i.e., the two letters are the same, is higher if the letters are drawn from the same alphabet than from different alphabets.

The probability of choosing two letters the same from ciphertext (i.e., two as or two bs or two cs or … or two zs) would be

$$I = \frac{n_a}{n} \times \frac{n_a - 1}{n-1} + \frac{n_b}{n} \times \frac{n_b - 1}{n-1} + \frac{n_c}{n} \times \frac{n_c - 1}{n-1} + \ldots + \frac{n_z}{n} \times \frac{n_z - 1}{n-1}$$

This number is denoted *I* and called the index of coincidence of the ciphertext.

Because Friedman denoted this number by the Greek letter kappa κ, it is sometimes called the Kappa Test.

English plaintext

The frequencies of the letters in English are:

Letter	a	b	c	d	e	f	g	h	i	j	k	l	m
Frequency	.082	.015	.028	.043	.127	.022	.020	.061	.070	.002	.008	.040	.024
Letter	n	o	p	q	r	s	t	u	v	w	x	y	z
Frequency	.067	.075	.019	.001	.060	.063	.091	.028	.010	.023	.001	.020	.001

Beker and Piper, *Cipher Systems: The Protection of Communications*, Wiley.

So, for a text in plaintext English, the probability of "drawing" two letters that are the same is:

aa	or	bb	or	cc	or	…	or	zz
$.082\times.082$	+	$.015\times.015$	+	$.028\times.028$	+	…	+	$.001\times.001$

This probability of "drawing" two letters that are the same – the index of coincidence -- is approximately $I \approx 0.0656010$.

Monoalphabetic Ciphers

If the ciphertext were generated by a monoalphabetic cipher, we should determine *I* to be near 0.065 because a monoalphabetic cipher is just a permutation of the letters of a single alphabet. The frequencies of letters for the ciphertext alphabet should be nearly the same as for English – but in a different order.

Polyalphabetic ciphers

If more than one alphabet were used, the frequencies of the letters should be more nearly uniform. In the ideal case, the frequencies of ciphertext letters would be uniform. If the frequencies were uniform, the probability of "drawing" two letters that were the same would be:

$$I \approx \underbrace{\left(\frac{1}{26}\times\frac{1}{26}\right)+\left(\frac{1}{26}\times\frac{1}{26}\right)+\left(\frac{1}{26}\times\frac{1}{26}\right)+\ldots+\left(\frac{1}{26}\times\frac{1}{26}\right)}_{\text{26 terms}}=\frac{1}{26}\approx 0.038.$$

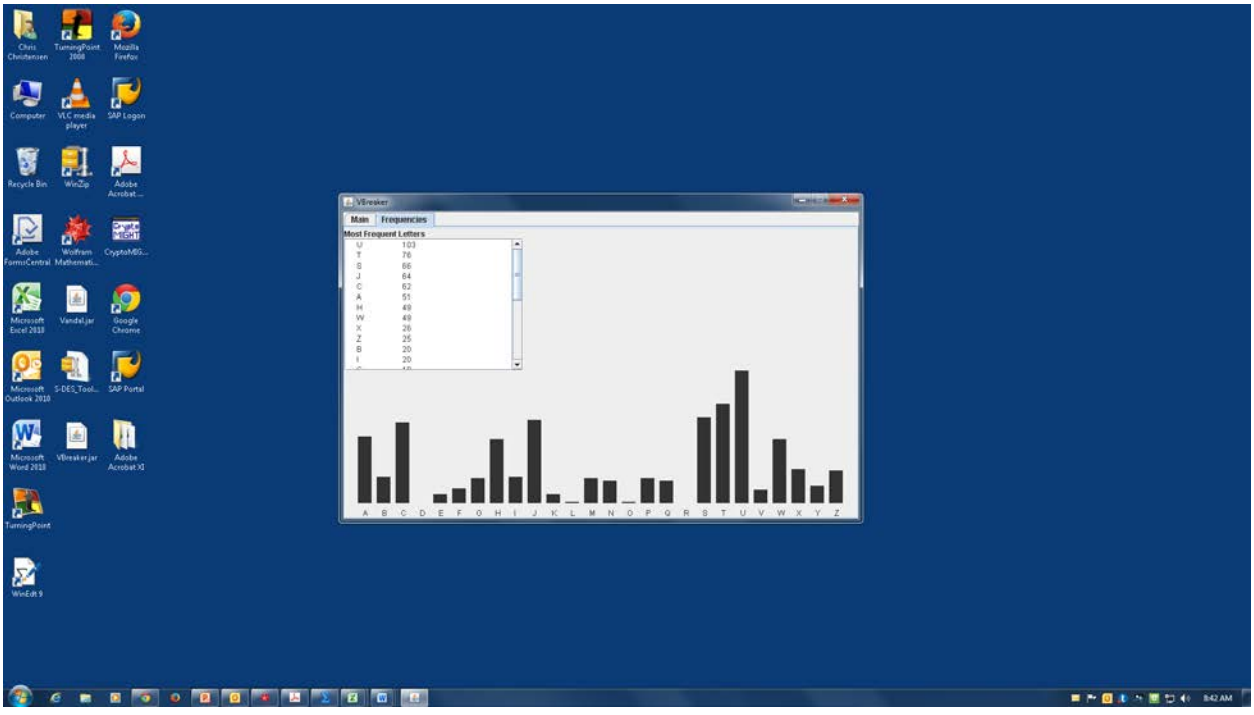
For polyalphabetic ciphers, the frequencies of the letters would become more nearly uniform – more nearly the same for each letter. We should determine *I* to be near 0.038.

Determining Whether a Cipher is Monoalphabetic or Polyalphabetic

Recall that, using frequency analysis, peaks and valleys of frequencies suggest a monoalphabetic cipher and relatively uniform frequencies suggest a polyalphabetic cipher. The Friedman test is a statistical way of "looking for peaks and valleys versus uniform frequencies."

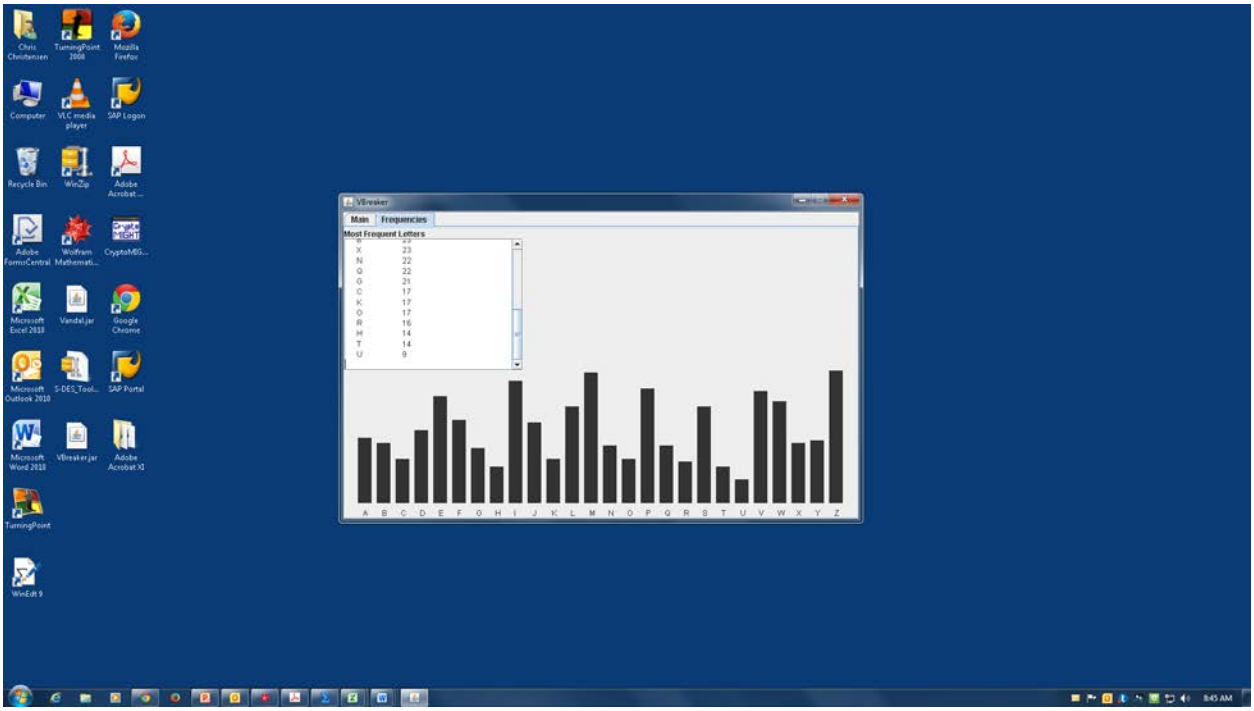
We test the ciphertext by calculating *I* based on the ciphertext frequencies. The closer that *I* is to 0.065, the more likely it is that we have a monoalphabetic cipher. The closer that *I* is to 0.038, the more likely that we have a polyalphabetic cipher.

The following frequencies are for ciphertext enciphered with an affine cipher:



The peaks and valleys of frequencies suggest a monoalphabetic cipher. The index of coincidence is 0.0701.

The following frequencies are for ciphertext enciphered with a Vigenère cipher:



The relatively uniform frequencies suggest a polyalphabetic cipher. The index of coincidence is 0.0441.

Estimating the Length of the Keyword

William Friedman's index of coincidence can also be used to estimate *l* the length of the keyword of a Vigenère cipher.

We will develop an approximation formula for *I*, the index of coincidence; this formula will contain *l* and *n*, the number of letters in the ciphertext. Then, to get an approximation for the length *l*, we will solve for *l* in terms of *I* and *n* (we know *n* and can calculate *I*).

First, assume that we know *l* and arrange the ciphertext into *l* columns. Now each column corresponds to a Caesar cipher. Although the columns might not all have the same length, we will assume that the number of letters in the ciphertext is large enough so that we can assume that they each have length $\frac{n}{l}$; i.e., we will assume that the error using this number for the length of each column is not large.

If we chose two letters from the ciphertext, what is the probability that they come from the same column and are the same letter?

One possibility is that we select two letters from the ciphertext that come from the same column and are the same letter. What is that probability?

Select a letter from the ciphertext. This selection determines a column. The probability that the next letter chosen comes from the same column is $\frac{\frac{n}{l}-1}{n-1}$. Because both letters are selected from the same Caesar cipher alphabet, the probability that both are the same is approximately the same as for standard English 0.065. So, the probability that both letters are selected from the same column and are the same letter is approximately $\frac{\frac{n}{l}-1}{n-1}\times 0.065$.

The other possibility is that we select two letters from the ciphertext that come from different columns but are the same letter. What is that probability?

Select a letter from the ciphertext. Again, this determines a column. The probability that the next letter comes from a different column is $\frac{n-\frac{n}{l}}{n-1}$. Because the two letters are selected from different Caesar cipher alphabets, the probability that both are the same is approximately the same as for a random alphabet 0.038. So, the probability that both letters are selected from different columns and are the same letter is approximately $\frac{n-\frac{n}{l}}{n-1}\times 0.038$.

So we have two cases: the two letters are selected from the same column and are the same letter or the two letters are selected from different columns and are the same letter. To get an approximation of the index of coincidence *I*, the probability that the two letters selected are the same, we add these two probabilities:

$$I \approx \frac{\frac{n}{l}-1}{n-1} \times 0.065 + \frac{n-\frac{n}{l}}{n-1} \times 0.038 .$$

Doing a bit of algebra to solve for *l*, we obtain:

$$I \approx \frac{\frac{n}{l}-1}{n-1} \times 0.065 + \frac{n-\frac{n}{l}}{n-1} \times 0.038$$

$$(n-1)I \approx \left(\frac{n}{l}-1\right) \times 0.065 + \left(n-\frac{n}{l}\right) \times 0.038$$

$$(n-1)I \approx \frac{n}{l} \times 0.065 - 0.065 + n \times 0.038 - \frac{n}{l} \times 0.038$$

$$(n-1)I + 0.065 - 0.038n \approx \frac{n}{l} \times (0.065 - 0.038)$$

$$(n-1)I + 0.065 - 0.038n \approx 0.027\frac{n}{l}$$

$$l \approx \frac{0.027n}{(n-1)I + 0.065 - 0.038n}$$

A commonly used table to estimate the length of the keyword is:

Estimated length of keyword	Index of Coincidence
1	0.0660
2	0.0520
3	0.0473
4	0.0449
5	0.0435
6	0.0426
7	0.0419
8	0.0414
9	0.0410
10	0.0407
⋮	⋮
∞	0.0388